Mathematical Formalism

by James M. Hyman and E. Ann Stanley

We will build up the equations for our risk-based model of AIDS through successive modifications of the basic equation of epidemiology, the equation of mass action. Its simplest form is given by

$$\frac{dI}{dt} = \alpha I \left(1 - \frac{I}{N}\right), \tag{1}$$

where $I(t)$ is the number infected, N is the total population and α is a constant. Equation 1 describes the spread of HIV infection by random sexual contact among a sexually active population of fixed size N. As explained in the main text, if a population mixes homogeneously, this equation gives rise to an initial exponential growth in the number infected with constant relative growth rate of α.

As the number infected becomes comparable to the total population the growth rate will decrease, so we rewrite Eq. 1 to show that time dependence:

$$\frac{dI}{dt} = \lambda(t)S(t), \tag{2}$$

where $S(t) = N - I(t)$ is the number of persons susceptible to infection and $\lambda(t) = \alpha I(t)/N$. So far the only independent variable is time t and $\lambda(t)$ is the time-dependent relative growth rate of the number infected.

To describe the AIDS epidemic over long times, we must account for individuals who eventually develop AIDS and die. Thus the total population will not remain constant but will change with time. We divide the population into three sectors: the sexually active, uninfected susceptibles $S(t)$; those infected with HIV who do not have AIDS $I(t)$; and people with AIDS $A(t)$. We assume the susceptibles and the infected are sexually active (and therefore can infect others) but that those with AIDS are not. Thus the sexually active population $N(t)$ is equal to $S(t)+I(t)$. Moreover, we assume that people mature, or migrate, into the sexually active susceptible population and retire from it at a constant relative rate μ, so that in the absence of AIDS the susceptible population would remain constant at the value S_0, that is, $N(t) = S(t) = S_0$ in the absence of HIV.

We also introduce the parameter γ, the relative rate at which people who are infected develop AIDS, and δ, the relative rate at which people die from AIDS.

Now we can write down a set of rate equations for changes in $S(t), I(t)$ and $A(t)$ with time.

The rate of change in the number infected is like Eq. 2 except the right-hand side includes negative terms that account for decreases due to conversion to AIDS at a rate $\gamma I(t)$ and aging of the infected at a rate $\mu I(t)$:

$$\frac{dI(t)}{dt} = \lambda(t)S(t) - (\gamma + \mu)I(t). \tag{3}$$

The number of uninfected susceptibles increases through maturation of "juveniles" at a rate μS_0 and decreases through aging at a rate $\mu S(t)$ and through infection with HIV at a rate $\lambda(t)S(t)$:

$$\frac{dS(t)}{dt} = \mu(S_0 - S(t)) - \lambda(t)S(t). \tag{4}$$

The number of people with AIDS increases through conversion of infecteds at a rate $\gamma I(t)$ and decreases through death at a rate $\delta A(t)$:

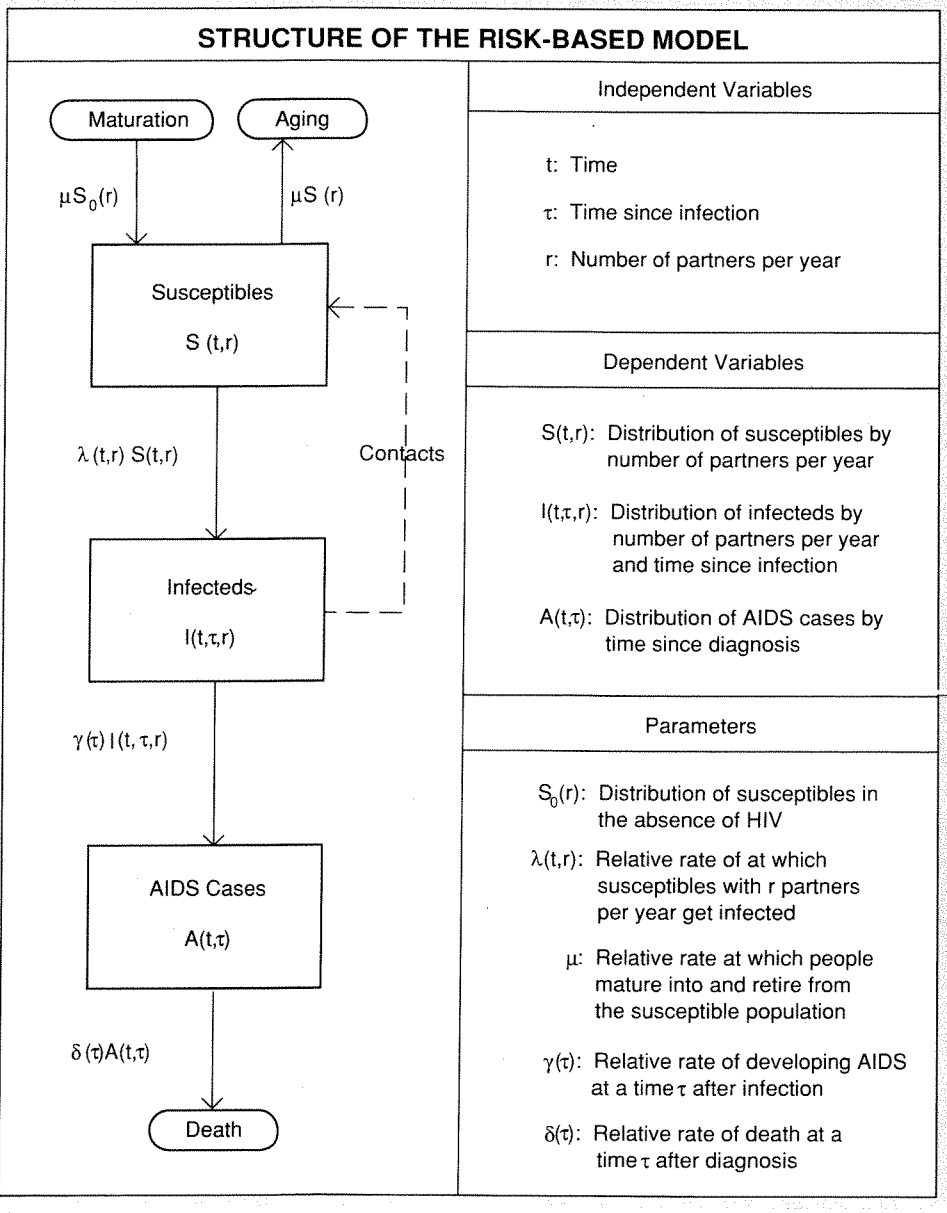
$$\frac{dA(t)}{dt} = \gamma I(t) - \delta A(t). \tag{5}$$

The accompanying block diagram illustrates the inputs and outputs to each of the three sectors of the population.

The most important assumptions in any model of AIDS are embedded in the definition of $\lambda(t)$, the rate of infection per susceptible. In the simple model just presented, all members of the population are assumed to be equal in their susceptibility and the rate of infection per susceptible is given by

$$\lambda(t) = i\ c\ p\ \frac{I(t)}{I(t)+S(t)}, \tag{6}$$

where the constant i is the probability of infection per sexual contact, the constant c is the average number of sexual contacts per partner, the constant p is the average number of partners per year, and $\frac{I(t)}{I(t)+S(t)}$ is the infected fraction of the sexually active population.



Note that this simple model produces exponential growth at the start of the epidemic. All members are equally at risk (homogeneous mixing) and the probability of infection per contact i remains constant throughout the years of infection.

We will now modify the simple model defined by Eqs. 3–6 to account for two crucial aspects of the AIDS epidemic. First, since AIDS takes many years to develop and the infectivity during the period of infection may vary in time, we introduce an additional independent variable τ, the time since infection. Second, since individuals who are very active sexually and who change partners frequently have a greater risk of becoming infected, we introduce the variable r, which quantifies the level of risky behavior in the sexually active population. In this model, r is defined as the number of new partners per year.

Using the two new independent variables τ and r, we distribute $I(t), S(t)$ and $A(t)$ over risk behavior and/or time since infection. (See the definitions in the block diagram.) In addition, the constant S_0 is the integral of an equilibrium distribution over risk behavior, $S_0 = \int_0^\infty S_0(r)dr$. Note that $S_0(r)$ corresponds to $N(r)$ in the main text; also the main text presents evidence that $S_0(r) \propto r^{-3}$ for large r.

We can now write down the equations of our risk-based model that correspond to Eqs. 3–5. Equation 3 for the infected population is replaced by Eqs. 7a and b. Equation 7a specifies that the rate at which people of risk r are becoming infected is $\lambda(t,r)S(t,r)$. Equation 7b says that rate at which the infecteds develop AIDS is proportional to the conditional probability $\gamma(\tau)$, which is a function of the time since infection, and the rate at which they leave the population is proportional to μ.

$$I(t,0,r) = \lambda(t,r)S(t,r). \tag{7a}$$

$$\frac{\partial I(t,\tau,r)}{\partial t} + \frac{\partial I(t,\tau,r)}{\partial \tau} = -\gamma(\tau)I(t,\tau,r) - \mu I(t,\tau,r). \tag{7b}$$

Equation 8 for the susceptibles has a structure similar to that of Eq. 4 except that now the rate of infection per susceptible $\lambda(t,r)$ depends on the risk behavior r:

$$\frac{\partial S(t,r)}{\partial t} = \mu(S_0(r) - S(t,r)) - \lambda(t,r)S(t,r). \tag{8}$$

Equation 9a says that the rate at which AIDS cases are being diagnosed at time t is equal to the rate at which infecteds convert to AIDS, $\gamma(\tau)I(t,\tau,r)$, integrated over all risk behaviors r and times since infection τ. Equation 9b accounts for loss of AIDS cases due to death.

$$A(t,0) = \int_0^\infty \int_0^\infty \gamma(\tau)I(t,\tau,r)d\tau dr. \tag{9a}$$

$$\frac{\partial A(t,\tau)}{\partial t} + \frac{\partial A(t,\tau)}{\partial \tau} = -\delta(\tau)A(t,\tau). \tag{9b}$$

The major change in this new set of equations is the form we assume for $\lambda(t,r)$, the relative rate at which susceptibles with r partners per year get infected. We generalize Eq. 6 to include variation in the degree of sexual contact between individuals with different risk behaviors as well as variation in infectiousness with time since infection. The general form of $\lambda(t,r)$ is given by

$$\lambda(t,r) = r \int_0^\infty \int_0^\infty c(r,s)\rho(t,r,s)i(\tau)\frac{I(t,\tau,s)}{N(t,s)}d\tau ds, \tag{10}$$

where $c(r,s)$ is the average number of sexual contacts in a partnership between a person with risk r and one with risk s, $i(\tau)$ is the infectiousness at τ years since in-

fection, $\frac{I(t,\tau,s)}{N(t,s)}$ is the probability that a person with risk s will be infected at time τ, and $\rho(t,r,s)$ is the fraction of the partners of a person with risk r who have risk s. The total number of sexually active people with risk s is given by $N(t,s) = S(t,s) + \int_0^\infty I(t,\tau,s)d\tau$.

Equations 7–10 describe the basic structure of our risk-based model. It differs from the well-known model of Anderson and May in one major respect—the form of $\lambda(t,r)$. Anderson and May assumed homogeneous mixing among the entire population, that is, that partners are chosen purely on the basis of availability. Then $\rho(t,r,s)$, the fraction of the partners of a person with risk r who have risk s, is just the proportionate mixing value:

$$\rho(t,r,s) = \frac{sN(t,s)}{\int_0^\infty xN(t,x)dx}. \tag{11}$$

They also assumed that the average number of sexual contacts per partner and the infectiousness were constant, so that $\lambda(t,r)$ becomes

$$\lambda(t,r) = \frac{icr\int sI(t,s)ds}{\int xN(t,x)dx}. \tag{12}$$

This form for $\lambda(t,r)$ (adapted from the model of Hethcote and Yorke for the spread of gonorrhea) yields exponential growth for the early stages of the epidemic.

We suggest that the assumption of homogeneous mixing is sociologically unrealistic. Instead, we build into our model a general form for $\rho(t,r,s)$ that allows for biased mixing among the population. That is, $\rho(t,r,s)$ includes an acceptance function, $f(r,s)$, that specifies the frequency at which an individual with risk behavior r chooses a partner with risk behavior s. When the acceptance function $f(r,s)$ is 1, we return to homogeneous mixing. When $f(r,s)$ is a narrow Gaussian, for example, $f(r,s) = \exp\left(-(r-s)^2/\epsilon(r+a)^2\right)$, people choose partners who are similar to themselves. This latter assumption is presented in the main text and yields the power-law growth in AIDS cases seen in the data.

For completeness we give the general form of $\rho(t,r,s)$:

$$\rho(t,r,s) = \begin{cases} (1-\int_0^r \rho(t,r,x)dx)\frac{f(r,s)sN(t,s)}{\int_r^\infty f(r,x)xN(t,x)dx}, & \text{for } r \le s \\ \rho(t,s,r)\frac{sN(t,s)}{rN(t,r)}, & \text{for } r > s. \end{cases} \tag{13}$$

This complicated function satisfies three necessary properties:

1. The number of partners with risk behavior s chosen by people with risk behavior r is equal to the number of partners with risk behavior r chosen by people with risk behavior s; that is,

$$rN(t,r)\rho(t,r,s) = sN(t,s)\rho(t,s,r). \tag{14}$$

2. People with risk behavior r have r partners per unit time; that is,

$$1 = \int_0^\infty \rho(t,r,s)ds. \tag{15}$$

3. The fractions $\rho(t,r,s)$ are positive.

In order to study the effects of different mixing patterns on the growth of the epidemic, we have chosen various forms for the acceptance function $f(r,s)$ and then solved Eqs. 7–9 numerically. The results are presented in "Numerical Results of the Risk-Based Model of AIDS." Also presented there are numerical solutions for different assumptions about infectiousness from time since infection. ■

Numerical Results of the Risk-Based Model

by James M. Hyman, E. Ann Stanley, and Stirling A. Colgate

Here we will present numerical solutions to the full risk-based biased-mixing model. These solutions validate the simplified version of the model presented in the main text and illustrate how variations in the input parameters affect the predicted course of the epidemic. The equations and parameters of the model are defined in "Mathematical Formalism for the Risk-Based Model of AIDS," hereafter referred to as "Math Formalism." The model tracks the time evolution of three sectors of the population: the sexually active susceptibles $S(t,r)$; the sexually active infecteds $I(t,\tau,r)$; and the people with AIDS $A(t,\tau,r)$. It takes into account deaths due to AIDS and the long time between HIV infection and conversion to AIDS. It also allows us to vary assumptions about the infectiousness as a function of time since infection and the mixing between various risk groups in the population.

First we will assess the validity of the predictions in the main text. The analytic calculation presented there predicted that biased mixing among the sexually active population gives rise to a saturation wave of infection, which yields power-law growth in both the number infected and the number of people with AIDS. That calculation was based on the following assumptions: the initial susceptible population $S_0(r)$ is distributed in risk behavior as r^{-3} for r greater than the mean value of r; the infectiousness i is constant; the cumulative probability of conversion to AIDS $C(\tau)$ is zero for the first two years after infection and then increases linearly with τ at a rate such that every infected individual develops AIDS by 18 years after infection; and finally, the same fraction is infected in all risk groups before the start of the saturation wave. The wave of infection was then calculated as if each risk group had a growth rate proportional to r and grew to saturation independently of all other groups. That is, we did not account for mixing between people with different risk behavior because the calculation is too difficult to perform analytically. Moreover, AIDS cases and deaths were not removed from the infected population. The result was that the number infected grows as t^2 and the number of people with AIDS grows at t^3.

To check whether mixing among individuals with different risk behavior alters that result, we solved the full set of equations given in "Math Formalism." We used the same assumptions and conditions outlined above except that we allowed mixing between people with different risk behavior r. We found that when mixing is restricted to people whose risk behaviors are within a factor of 2 of each other, that is, the mixing is biased, a saturation wave of infection moves from high- to low-risk groups and the number infected grows as t^2, as predicted by the analytic calculation in the main text. Also, when mixing is random, or homogeneous, that is, is based only on availability, the number infected grows exponentially, the relative growth rate is constant, and the fastest growth occurs in the population with the most likely risk. Thus, doubling times for biased mixing are shorter initially and later become longer than those for random mixing.

Now let's consider numerical solutions to the full model under more general assumptions. We will first comment on their overall behavior and then present specific solutions. The numer-

THE RATE OF INFECTION $\lambda(r,t)$

The heart of the risk-based model is the complicated functional form of the rate of infection per susceptible with risk r, $\lambda(r,t)$ (see Eqs. 10 and 13 in "Math Formalism"). We will describe this function in words:

Rate of infection for a susceptible = Number of new partners per year × Rate of sexual contact between persons with risk behaviors r and s × Infectiousness per contact × Probability that a person with risk s is infected

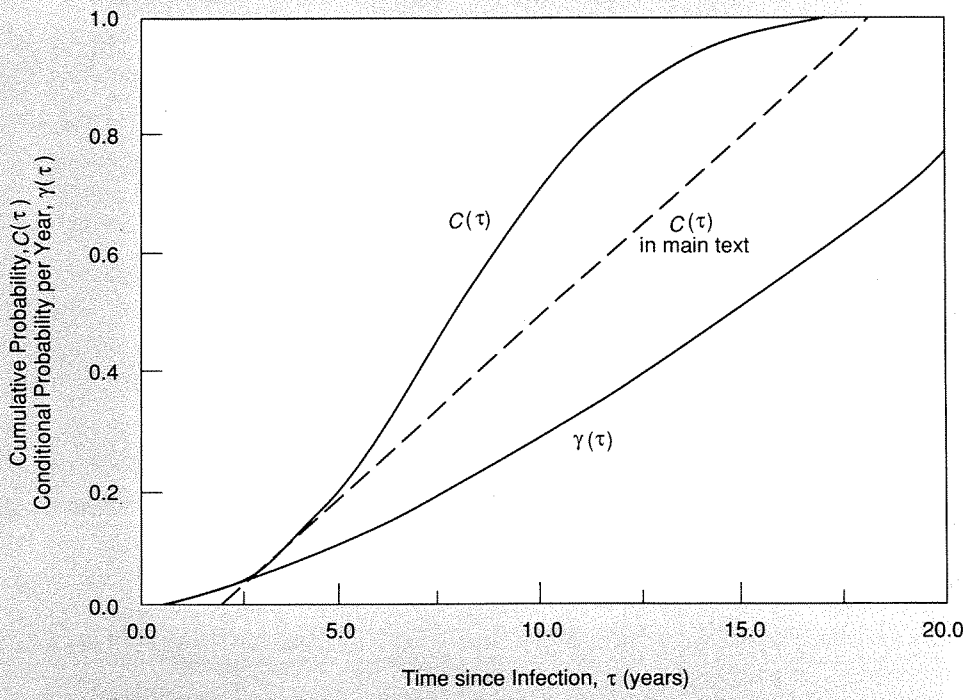
$$\lambda(r,t) = r\int_0^\infty \int_0^\infty \underbrace{c(r,s)}_{\text{Average number of sexual contacts in a partnership between persons with risk behaviors } r \text{ and } s} \underbrace{\rho(t,r,s;f(r,s))}_{\text{Fraction of partners of a person with risk behavior } r \text{ and who have risk behavior } s} i(\tau)\frac{I(t,\tau,s)}{N(t,s)}\,d\tau\,ds$$

The function $\rho(t,r,s)$ describes the level of mixing between people with risk behaviors r and s. It is defined in terms of an acceptance function $f(r,s)$ that determines the range from which partners are chosen.

ical results of the model change as we vary the input parameters $S_0(r)$, $I(t,0,r)$, $i(\tau)$, $A(t,0,r)$, $c(r,s)$, $f(r,s)$, $\gamma(\tau)$, μ, and $\delta(\tau)$ (see Fig. 1 in "Math Formalism" for the definitions of these parameters). The most critical parameters for determining the course of the epidemic are the initial distribution of risk behavior among the susceptible population $S_0(r)$ and the functions $i(\tau)$, $c(r,s)$, and $f(r,s)$, which determine the rate of infection per susceptible $\lambda(r,t)$ (see "The Rate of Infection"). In particular, the acceptance function $f(r,s)$ specifies the amount of mixing between different risk groups. Provided the mixing is biased, $S_0(r)$ decays as r^{-3} or r^{-4} and the numerical value of the product $c(r,s)i(\tau)$ is between 0.025 and 0.001 (this last provision determines the time scale of the epidemic), numerical solutions of our model show that the infection travels as a saturation wave from high- to low-risk groups for approximately the first 20 years. During those years the cumulative number infected and the cumulative number of people with AIDS grow as polynomials in time, rather than as exponentials.

By varying the functional forms of $\gamma(\tau)$, the rate, or conditional probability, of developing AIDS, and $i(\tau)$, the infectiousness since time of infection, we can raise or lower the degree of the polynomial growth, but in all of our calculations with biased mixing, the growth remains polynomial after the initial transients.

With these general remarks as background, we present various numerical solutions to the model. To obtain these solutions, Eqs. 9–10 in "Math Formalism" were integrated numerically with an explicit Adams-Bashford-Moulton solution method to an accuracy of 10^{-6} per unit time. The dependences on τ and r were calculated on a uniform grid of between 71 and 201 mesh points, and the convergence of solutions has been verified to within a few per cent.



RATE OF CONVERSION TO AIDS

Fig. 1. The rate of conversion to AIDS at time τ after infection $\gamma(\tau)$ is equal to the conditional probability that a person who did not have AIDS before time τ develops AIDS at time τ. Thus, it is given by $\gamma(\tau) = \frac{dC(\tau)/d\tau}{1-C(\tau)}$, where $C(\tau)$ is the cumulative probability of developing AIDS at τ years after infection. The figure shows plots of the functions $\gamma(\tau)$ and $C(\tau)$ used in all the numerical solutions presented here. For comparison we also show a plot of the form for $C(\tau)$ assumed in the main text (dashed line).

We emphasize, however, that although the solution techniques are accurate, the equations are still crude approximations and the results are meant to illustrate the general behavior of the model, not to give accurate forecasts of the future. Even the full model is much too simplistic to be used as a predictive tool.

For all the solutions presented here, we assume an initial population of 10 million people whose risk behavior (which we identify as the number of new partners per year) is distributed as an inverse cubic with a mean of 24 partners per year. We use the initial distribution $S_0(r) = 20(1 + \frac{r}{24})^{-3}$. We also use that form of $\gamma(\tau)$, the conditional probability for converting to AIDS, shown in Fig. 1. (The relationship between $\gamma(\tau)$ and $C(\tau)$ is described in the figure caption.) We use the constant value $\mu = 0.02$ per year for the fractional rate of maturation. The fractional rate of deaths due to AIDS $\delta(\tau)$ is obtained from CDC data. Also, for simplicity in this series of calculations, we assume the number of contacts per partner $c(r,s)$ is a constant $\bar{c}$.

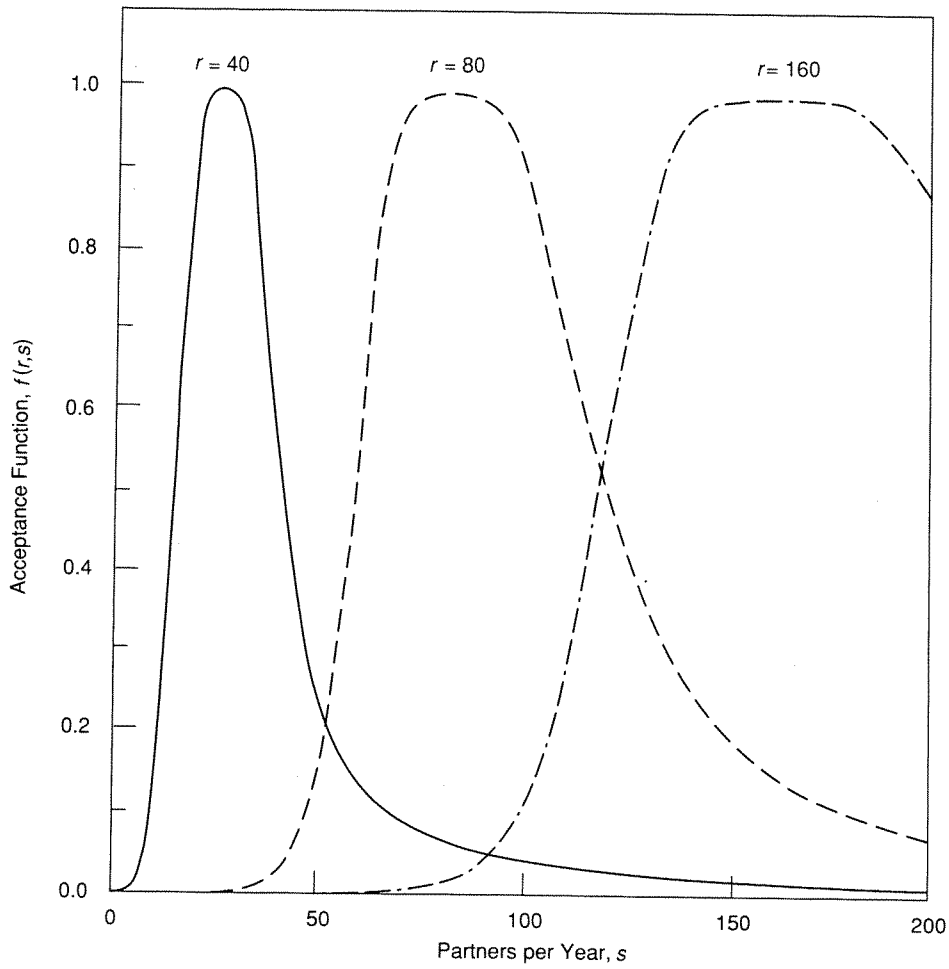
The parameter that we vary from one solution to another is $\lambda(r,t)$, the relative growth rate of infection among susceptibles with r partners per year. In particular we vary two factors in $\lambda(r,t)$: the acceptance function $f(r,s)$ and the in-

BIASED MIXING FOR BASELINE SOLUTION

Fig. 2. The numerical solutions presented here use an inverse quartic function for the acceptance function $f(r, s)$:

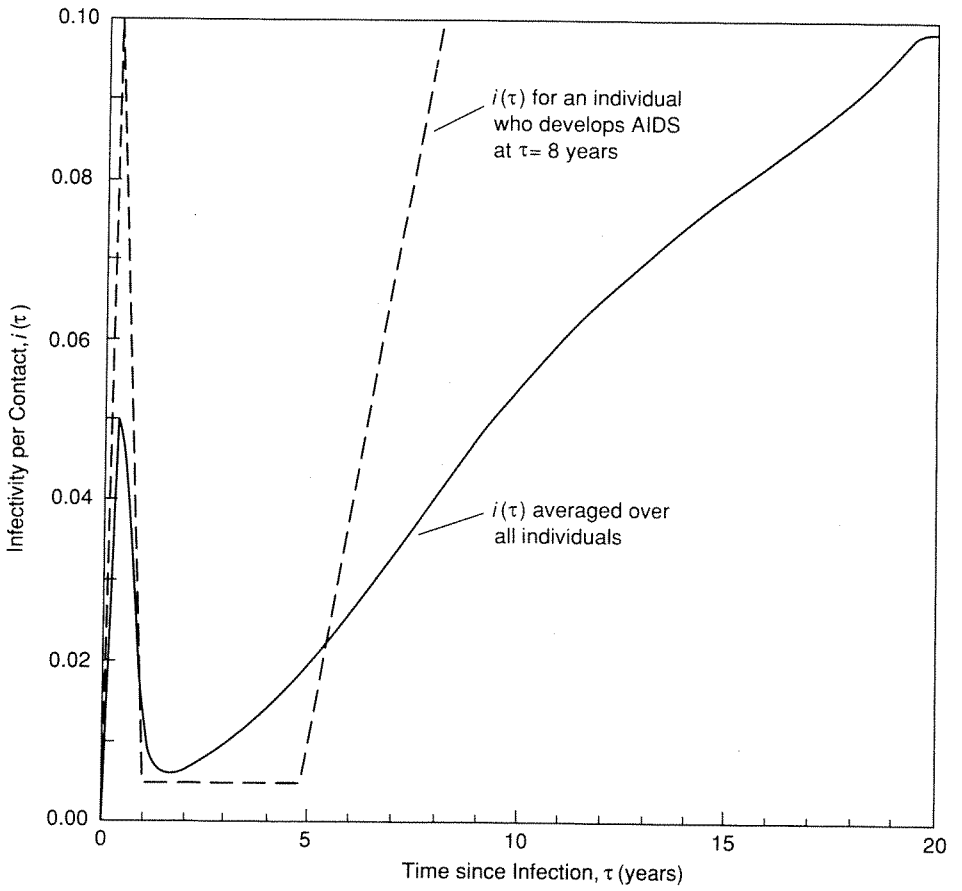
$$f(r, s) = \left[1 + \frac{(r - s)^4}{\epsilon(r + r_m)^4}\right]^{-1}.$$

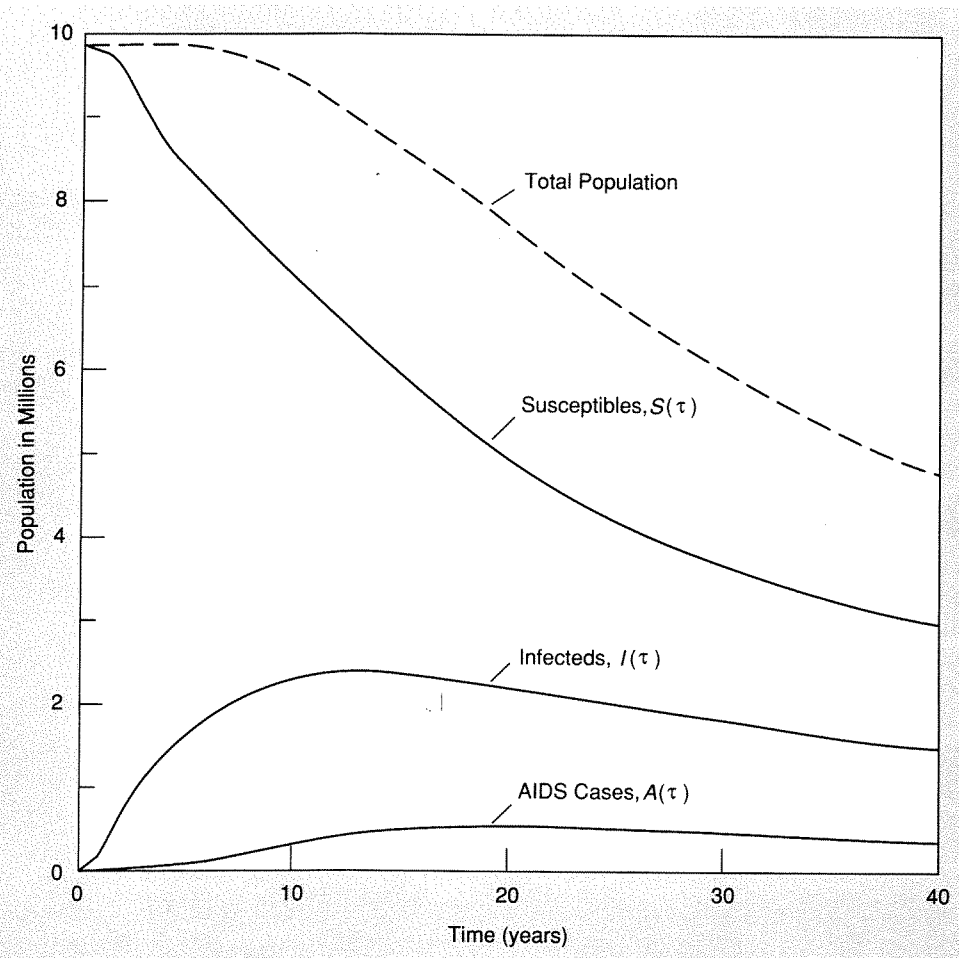
The figure shows $f(r, s)$ versus s for r = 40, 80, and 150 when ϵ = 0.01. For each value of r, $f(r, s)$ determines the fraction of partners with risk s chosen by people with risk r. Here $f(r, s)$ specifies that most partners of a person with risk r have risk behaviors between $\frac{1}{2}r$ and r; that is, the mixing is heavily biased toward people with similar risk behavior.



TIME-DEPENDENT INFECTIVITY

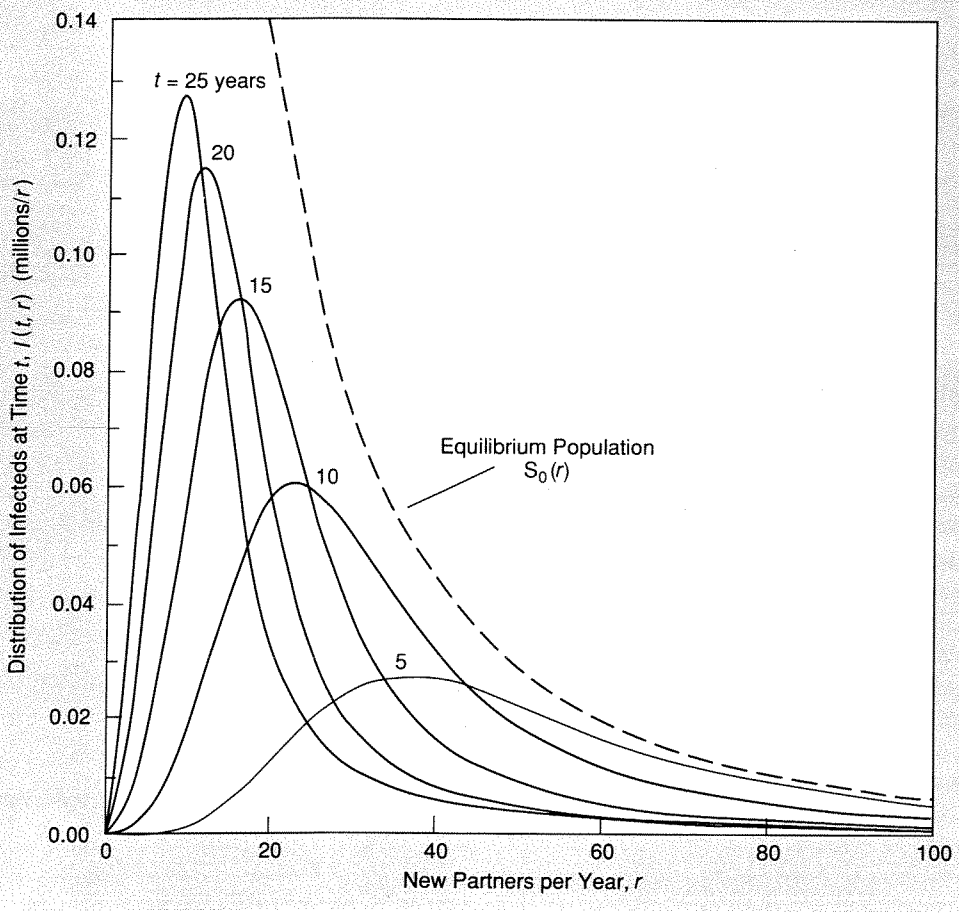
Fig. 3. The mean infectiousness $i(\tau)$ versus time since infection (solid line) used in all but the last solution presented here. The function $i(\tau)$ is an average over individuals each of whom develops AIDS at some time between 2 and 20 years since infection. The average infectiousness of each individual over the time from infection to AIDS is 0.025. The dotted line shows the pattern of infectiousness that we postulate for a single individual. In this case the individual develops AIDS 8 years after infection. The initial peak of infectiousness for this individual is always taken to be greater than 6 months because our numerical techniques are not yet designed to handle sharper peaks.





BASELINE SOLUTION

Fig. 4. The time-dependent behavior of various sectors of the population predicted by the baseline calculation. Despite a slow migration of people into the total population, the high mean new-partner rate of 24 partners per year drives an epidemic that substantially depletes the total population as a large fraction become infected and then die of AIDS. The very slow progression from infection to AIDS and rapid death from AIDS produces a delay between start of infection and the AIDS epidemic. Also, at all times many fewer people have AIDS than are infected.



SATURATION WAVE IN BASELINE SOLUTION

Fig. 5. Distributions of the number infected over number of new partners per year at times t = 5, 10, ..., 40 years during the baseline calculation. The dotted line shows the distribution of the total population in the absence of HIV. As time progresses, a wave of infection moves from high-risk to low-risk groups. Essentially all members of high-risk groups become infected, and the populations of those groups decrease to very low levels as everyone develops AIDS and dies. As the wave moves progressively through lower-risk groups, an ever smaller fraction of those groups becomes infected.

fectiousness per contact since time from infection $i(\tau)$.

We present first a "baseline" solution. The acceptance function $f(r,s)$ and the infectiousness per contact $i(\tau)$ for this solution are described in Figs. 2 and 3, respectively. The acceptance function in Fig. 2 is an inverse quartic function of r and s, which describes the probability that a person with risk behavior r chooses a partner with risk behavior s:

$$f(r,s) = \left[1 + \frac{(r-s)^4}{\epsilon(r+r_m)^4}\right]^{-1},$$

where $\epsilon = 0.01$ and $r_m = 10$ partners per year. The figure shows $f(r,s)$ versus s for three different values of r. As r increases, the width of the acceptance function increases. In rough terms, this function describes a biased mixing pattern in which a person with risk r chooses most of his or her partners from a group that ranges in risk behavior from $\frac{1}{2}r$ to $2r$.

Figure 3 is a plot of $i(\tau)$, the mean infectiousness per partnership versus time since infection. The mean infectiousness is an average over the infectiousness of many individuals each of whom develops AIDS at different times (determined by $\gamma(\tau)$) since the time of infection. Figure 3 also shows the infectiousness curve for an individual who develops AIDS 8 years after infection. The infectiousness for this individual is assumed to have an initial peak, a latency period of about four years, and finally a steady rise. The average infectiousness for each individual is assumed to be 0.025. The initial peak is about 6 months wide, probably too wide to be realistic, but our numerical code does not yet have the capability of resolving a burst that is only a few weeks in duration. Nevertheless, the wider shape that we have used serves the purpose of illustrating what the impact of an initial peak of infectiousness can be.

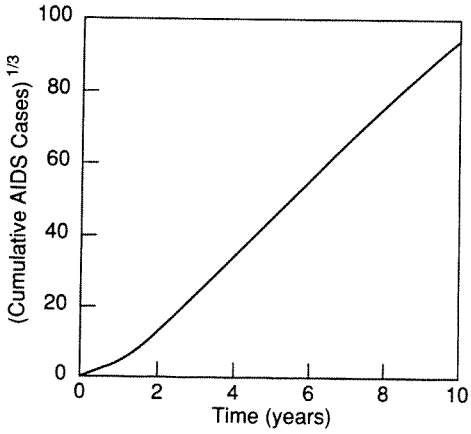
The infected population at $t = 0$ contains 1000 individuals distributed as a narrow Gaussian function of r centered at 175 partners per year and distributed linearly in τ. Although here we assume that the epidemic starts among the highest-risk groups, this choice does not have a major impact on the numerical results. In particular, if the infecteds at $t = 0$ are centered at the mean, the epidemic follows a similar course but starts about 2 years later. If the infecteds at $t = 0$ are distributed over all risk groups, the saturation wave takes off sometime between 0 and 2 years later.

The input parameters and initial conditions just described yield our "baseline" solution. Figure 4 shows $S(t), I(t)$, and $A(t)$ over a 40-year period. During that period about half of the population dies of AIDS. The number infected $I(t)$ and the number of people with AIDS at any given time $A(t)$ rise steadily for more than 10 years and then decline slightly as the epidemic reaches a steady state.

Figure 5 shows plots of the number infected versus risk behavior at times $t = 5$, 10, 15, 20 and 25 years. Here we see that the infection travels as a saturation wave from high- to low-risk groups. The wave takes 20 to 25 years to reach the lower-risk groups.

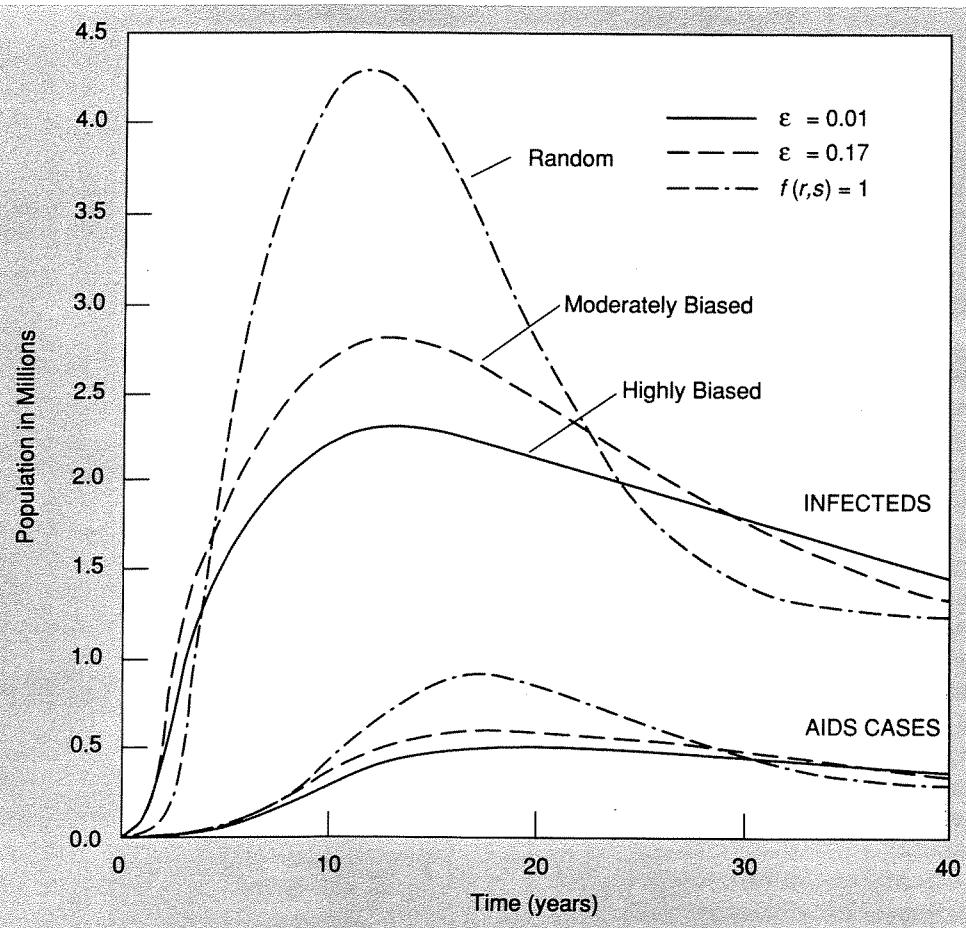
Figure 6 is a plot of the cube root of the cumulative number of AIDS cases as a function of time. The nearly straight line between 1 and 10 years shows that the calculation is not inconsistent with the observation that the number of AIDS cases grows as t^3 during the initial stages of the epidemic. The main reason that the growth is not purely cubic is the deviation of the initial profile $S_0(r)$ from a pure inverse cubic. However, the profile we chose for $S_0(r)$ fits the available partner-change-rate data much better than does Eq. 13 in the main text. We have also assumed a fairly large infectivity, which speeds up the progress of the entire epidemic. Consequently, by 10 years from the start of the saturation wave, the wave front has reached the lowest-risk populations, which, in turn, slow down the cubic growth. Although the solution just presented roughly matches the observed cubic growth of AIDS, it does not prove that the input parameters are correct but rather suggests the basic ingredients needed to produce the type of epidemic we are experiencing. A slightly different mix of input parameters yields very similar growth.

The assumption of biased mixing is the feature that sets this model apart from other models. Let's see how the epidemic changes when this assumption is relaxed. Figure 7 shows three solutions to the model that differ only in the



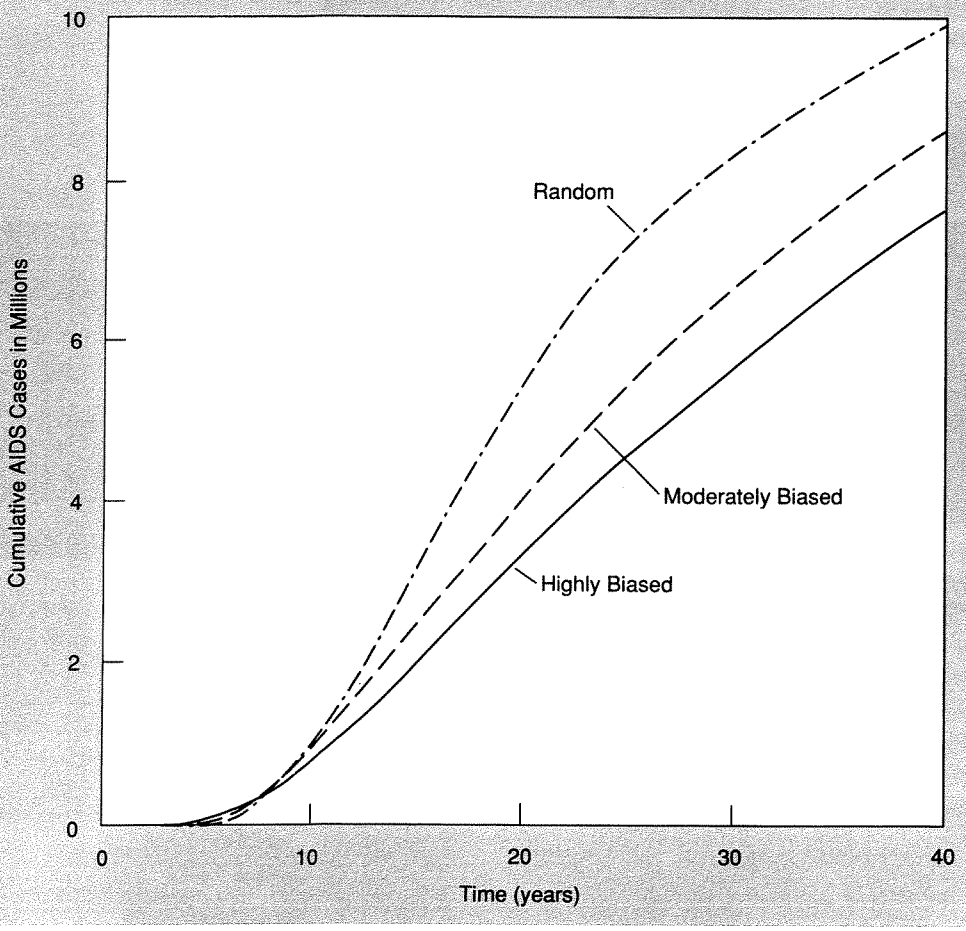
"CUBIC GROWTH" OF BASELINE SOLUTION

Fig. 6. The cube root of the cumulative number of AIDS cases as a function of time for the baseline solution. Although the curve is not perfectly straight, a t^3 growth in the cumulative number of AIDS cases is a good fit to this calculation between $t = 1$ and $t = 9$ years. Thus, despite the many complexities included in the numerical model, its solutions behave quite similarly to the analytic calculation of the main text. Note that the calculated time scales are fixed by the average value we assume for the product $c(r, s)I(\tau)$ and are therefore highly uncertain.



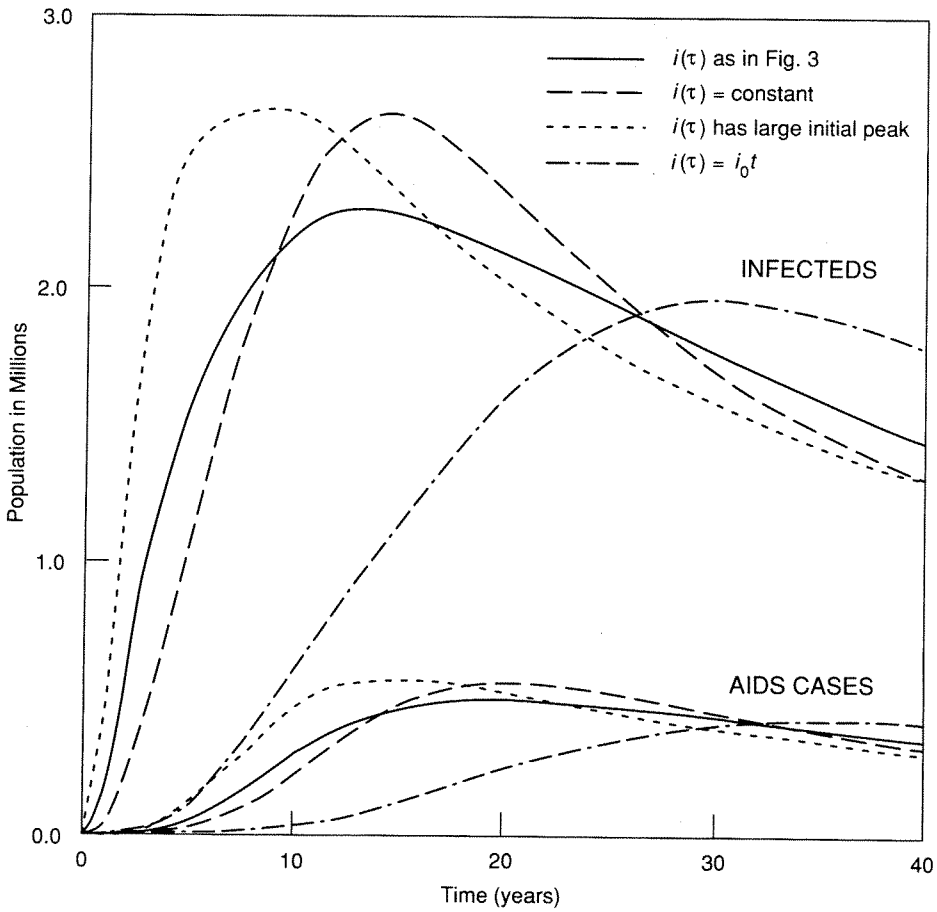
BIASED VERSUS RANDOM MIXING

Fig. 7. Time-dependent behavior of the number infected and the number of AIDS cases for various degrees of mixing among people with different risk behaviors. The baseline calculation (solid line) corresponds to the highest bias, or narrowest range of mixing. As the range of mixing widens, the epidemic changes dramatically. The growth pattern of the number infected appears to change more than that of the AIDS cases partly because of the scale of the plot, and partly because the slow conversion to AIDS smears out the effects of the change in the number infected. More biased mixing produces a more rapid initial growth than does random mixing, but growth slows down as the infection spreads among low-risk people and the total epidemic is smaller than that produced by random mixing. When mixing is random, high- and low-risk people, pass the virus back and forth between them, so an infected person is much more likely to encounter an uninfected person until the whole population saturates.



CUMULATIVE GROWTH IN AIDS AS MIXING VARIES

Fig. 8. Cumulative AIDS cases versus time for the calculation in Fig. 7. When mixing is random, the cumulative number of AIDS cases grows exponentially until the entire population reaches saturation of infections. When the mixing is highly biased, the number grows more as a polynomial.



EFFECTS OF VARYING THE INFECTIVITY

Fig. 9. The distribution of number infected $i(\tau)$ as a function of new-partner rate at t = 10 years for the calculations in Fig. 7. This figure demonstrates most dramatically the effects of varying the mixing patterns. When people have a strong bias to mix with others of similar risk, few people of low risk are infected in the early stages of the epidemic. In contrast, when partners are chosen purely on the basis of availability, people of low risk are infected early. The fact that early AIDS cases and early cases of infection were among people with high new-partner rates is evidence for biased mixing in the U.S. population.

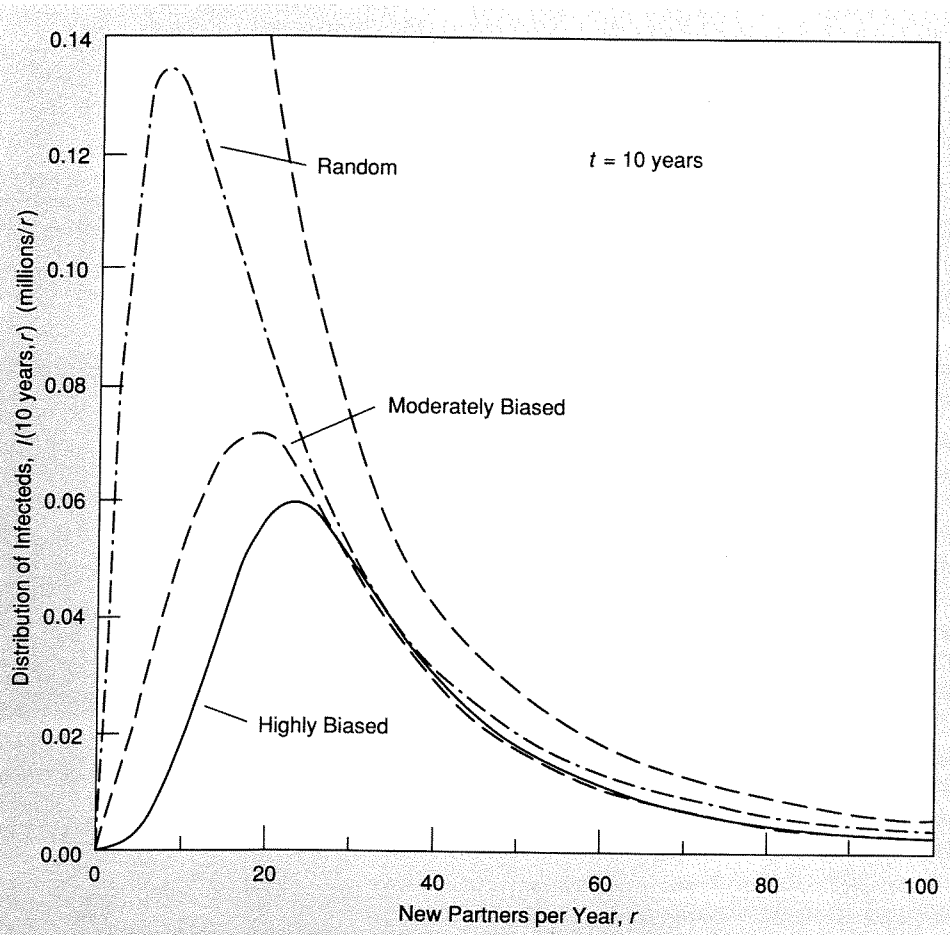
level of mixing among different risk groups. The solid lines show the baseline solution in which the mixing is strongly biased; that is, $f(r,s)$ is an inverse quartic with $\epsilon = 0.01$ (see Fig. 2). The dotted lines show a solution with less bias; that is $f(r,s)$ is again an inverse quartic but $\epsilon = 0.17$ so the curves of $f(r,s)$ versus s for different values of r have much wider peaks than those in Fig. 2. The dashed lines show a solution with no bias; that is, $f(r,s) = 1$ corresponding to random, or homogeneous, mixing. Note that as the mixing becomes less biased, the epidemic starts off slightly later but then grows faster because the doubling time increases at a slower rate.

Figure 8 shows the cumulative number of people with AIDS as a function of time for the three types of mixing. For random mixing, the number of people with AIDS grows nearly exponentially; that is, the doubling time is nearly constant. As the mixing becomes more biased, the number of people with AIDS grows more like a low-order polynomial.

It is worth cautioning that the initial distribution of infecteds, which is arbitrary, can have a significant impact on the early growth of the epidemic, especially if the initial growth rate is low. For the random-mixing case, growth in infections is so low initially that most people getting AIDS in the first 10 years were infected at $t = 0$. Consequently, since those infected at $t = 0$ were distributed linearly with τ, the number of AIDS cases grows as a polynomial during the first 10 years, and only the number infected grows exponentially. After 10 years both the number infected and the number of AIDS cases grow exponentially. For the cases of more-biased mixing, the initial growth in number of infecteds is more rapid, so the initial distribution $I(0,\tau)$ affects the solutions for a shorter period of time. Since our initial conditions are arbitrary, rather than based on knowledge of the earliest stages of the epidemic, the solution transients just described are also arbitrary.

Figure 9 shows the number infected versus risk behavior at $t = 10$ years for each of the three mixing patterns. We see that random mixing not only produces a faster-growing epidemic but also causes the epidemic to reach the low-risk groups almost immediately. Figures 4a and 4b of the main text also illustrate that point. The solution with biased mixing shows a saturation wave of infection traveling from high-to low-risk groups, but the solution with homogeneous mixing shows no such wave. Instead, the majority of those infected are always in the low-risk groups. Since the average partner rates for the earliest AIDS cases and infected homosexuals were high compared to the mean in the general homosexual population, these numerical results support the conclusion in the main text that biased mixing has produced the cubic growth of the AIDS epidemic.

We will now examine the effects of varying the function $i(\tau)$, the infectiousness since time of infection. In the main text we used a constant value of $i(\tau)$, but we also discussed the effects of a variable infectiousness. Here we display four solutions, each of which uses a different function for $i(\tau)$ (see Fig. 10). In all cases the mean infectiousness of an individual over the course of infection is 0.025. The solid lines correspond



INFECTEDS VERSUS RISK AS MIXING VARIES

Fig. 10. Time-dependent behavior of the number infected and the number of AIDS cases for various assumptions about the time-dependence of infectiousness. In these calculations we assign the same value for the average invectivity of any individual over the course of the epidemic and vary only the distribution of infectivity with time. A burst of infectivity just after infection causes the disease to spread very rapidly in the high-risk groups but has less effect as the disease spreads to groups with lower new-partner rates. A slowly rising infectivity several years after the initial burst sustains the epidemic in low-risk groups. With no initial burst of infectivity, but only a slow increase from infection until death, the epidemic initially spreads very slowly, but as more people approach the later stages of infection, the epidemic gains momentum. Without control measures the epidemic may eventually affect as many people as the other examples shown in the figure.

to the baseline solution shown earlier; $i(\tau)$ for that solution is shown in Fig. 3. The dashed lines are the solution when $i(\tau)$ is constant. The dotted lines are the solutions when the infectivity of a person getting AIDS at 8 years has a very large initial peak, then a 4-year period during which $i(\tau) = 0$, and finally a slow increase in $i(\tau)$ up to 8 years since infection. The dash-dot lines are the solutions when $i(\tau)$ has no initial peak, but instead, a person's infectivity increases continuously between the time of infection and the time of AIDS. A large initial peak in $i(\tau)$ produces the fastest-growing epidemic, the absence of an initial peak produces the slowest-growing epidemic, and a constant value for $i(\tau)$ produces an epidemic that is closest to the baseline solution but grows a bit more slowly at first, then somewhat faster, and finally approaches a similar steady state. (Note that the vertical scale in Fig.10 is a blow up of the vertical scale in Fig. 7.) In all cases the growth is "polynomial" in that the doubling times increase continuously. Nevertheless, the shape of $i(\tau)$ has a significant impact on the course of the epidemic.

Without better data for $i(\tau)$, the future course of the present epidemic cannot be estimated. Similarly, adequate data on the mixing patterns among different risk groups is sadly lacking. If nothing else, our risk-based model points out the areas for which more data are needed. We hope that this work will help to guide the data collection and analysis efforts that are now under way. ■

The Seeding Wave

by Stirling A. Colgate and James M. Hyman

Let us assume that our risk-based model is a reasonable description of how AIDS has grown since the time when a member of the highest risk group was infected. In other words, we assume the infection spread as a saturation wave from the highest risk group down through lower and lower risk groups. The question remains—what happened *before* the start of the saturation wave? Did an individual from the highest risk group become infected first and start the saturation wave immediately, or did an individual from a much lower risk (and therefore much larger) group start a slow spread of infection from lower to higher groups prior to the saturation wave? We call a slow spread of infection from lowest to highest risk groups a *seeding wave*. Now, if a seeding wave *were* started, do subsequent seeding events circumvent the slow spread by leapfrogging the infection to the highest risk group, thereby reducing the number infected before the start of the saturation wave?

Here we present a model of a seeding wave consistent with our saturation-wave model of subsequent growth. In particular, the model incorporates the same distribution of risk behavior and assumptions about biased mixing used in our risk-based model. We argue that, provided these assumptions are correct, the seeding wave is a likely scenario for the early spread of HIV infections in the United States. Moreover, the model predicts that the earliest HIV infection occurred in the mid-sixties, a prediction consistent with the first recognized case of AIDS in St. Louis in 1969.

Early Growth. Suppose the first infection in the United States is initiated, say, by either a visitor with HIV or a U.S. person visiting elsewhere. Although these two cases would not be equivalent if high risk of infection is correlated with high rate of travel, we will not consider such correlations here. Rather, we assume that risk of infection can be quantified using a single variable r with its distribution $N(r)$ defined by Eq. 13 in the main article. Since the probability of a person becoming infected is proportional to r, the probability $P(r)$ that at least one individual with risk r or greater becomes infected is given by

$$P(r) \propto \int_r^\infty N(r)rdr = r^{-1}, \qquad (1s)$$

for $r \geq 1$, that is, for the high-risk end of the population defined by Eq. 13.

Hence, the smaller r is, the greater is the probability that at least one individual of risk group r becomes infected despite the lower risk per individual. Also, the most likely case is that the first infected individual was a member of the average group, the group with $r = 1$.

A Simple Numerical Model. We wish to model the progression of the infection to higher risk groups starting with an infected individual close to the average. To help understand this process, we simplify by saying that the kth risk group r_k varies in risk behavior by a factor of 2, that is, r varies from r_k to $2r_k$. Hence, the various groups will have risk behaviors 1, 2, 4, 8... times the average. The number of individuals for $r > 1$ in the kth group (using Eq. 13) is

$$\int_{r_k}^{2r_k} N(r)dr = -\frac{1}{2}N_0 r^{-2}\Big|_{r_k}^{2r_k} = \frac{3}{8}N_0 r_k^{-2}. \qquad (2s)$$

Since the total population (the integral of Eq. 13 from $r = 0$ to $r = \infty$) is $\frac{3}{2}N_0$, our first group with $1 < r < 2$ is one-fourth of the total, and if we restrict ourselves to the homosexual population, one-fourth of the total is one million. Thus, the second group, with $2 < r < 4$, will have $(\frac{1}{4})(\frac{3}{8})N_0$ individuals, or $\frac{1}{16}$th of the total population, or 250,000. The third group, with $4 < r < 8$, will have $(\frac{1}{16})(\frac{3}{8})N_0$, or $\frac{1}{64}$th the total population, and so forth.

We do not believe that people exhibiting a preference for each other are likely to recognize a behavior difference much finer than a factor of 2, and hence, we use this rather crude measure of a group. We also suppose, as a reasonable but unknown example, that the fraction of the time an individual participates in risk outside his group is $F = \frac{1}{4}$. If this fraction is greater or less by a factor of 2, it will change what follows by a factor of 2, but that change is within the accuracy of these estimates.

We next ask how many individuals must be infected in group 1 before a member of group 2 is infected. There are one-fourth as many people in group 2 as in group 1 with twice the risk behavior; that is, the number of each group decreases as $1/r_k^2$ (from Eq. 2s), and they have contacts with other groups only one-fourth of the time. This fraction of out-of-group mixing will be distributed between both higher and lower risk groups.

Let us assume that the fraction is evenly divided between the higher and lower groups and, because of the same bias that leads to group preference, is primarily in the adjacent groups. Crudely then, F can be considered to be a diffusion coefficient. A bias towards only adjacent out-of-group mixing prevents the infection from leapfrogging to much higher groups and circumventing the slow seeding-wave progress.

The seeding wave progresses from group k to the next higher group $k + 1$ when one member of the next higher group is infected. If I_k is growing exponentially, $I_k = e^{(1-F)\alpha r t}$, then the cumulative probability of infecting one member of group $k + 1$, starting at the time when one member of group k is infected, is

$$\begin{aligned} I_{k+1} &= \int_0^t \frac{F}{2}(1-F)(\alpha r_k) I_k dt \\ &= \frac{F}{2} e^{(1-F)\alpha r_k t}\Big|_0^t \\ &= \frac{F}{2}(I_k - 1), \end{aligned} \tag{3s}$$

where the factor $\frac{F}{2}$ is needed because only one half of the out-of-group mixing pertains to the higher risk groups. The remaining half augments the growth rate of the next lower risk group.

Since $I_{k+1} = 1$ when the seeding wave progresses by one group, I_k at this transition becomes equal to $\frac{2}{F+1}$. Therefore, in our example (for which $F = \frac{1}{4}$), nine members of a group must become infected before a member of the next higher group becomes infected. The time for this to occur will be

$$t_k = \frac{\ln(\frac{2}{F} + 1)}{(1-F)\alpha r_k}. \tag{4s}$$

Thus, the speed of the seeding wave is $dk/dt = 1/t_k$. The remaining time for the seeding wave to go from group k to the highest risk group at $k = m$ is

$$t_{km} = \sum_k^m t_k, \tag{5s}$$

or

$$\begin{aligned} t_{km} &= \frac{\ln(\frac{2}{F} + 1)}{(1-F)\alpha} \sum_k^m \frac{1}{r_k} \\ &\simeq 2\frac{\ln(\frac{2}{F} + 1)}{(1-F)\alpha}, \end{aligned} \tag{6s}$$

for $m \gg k$. That is, the sum $1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} \cdots \simeq 2$ after even just a few terms. Thus, for most of the groups with $k < m$, the remaining time needed for the seeding wave to move through essentially all groups (that is, all but the few of highest risk) is just double the time to infect the adjacent next higher group. Now, each of these seeded groups is growing exponentially so that, as the time increases from, say, t_k to $2t_k$, the number infected in the kth group increases from I_k to I_k^2. Thus, the number of individuals infected in each group at the time the seeding wave ends, $t = t_m$, will be the square of the number infected when the next higher adjacent group is seeded with one individual, or $(\frac{2}{F} + 1)^2$. That is,

$$\begin{aligned} I_k(t_m) &= \int_0^{t_m} e^{(1-F)\alpha t} dt \\ &\simeq (\frac{2}{F} + 1)^2 \quad \text{for } k \ll m. \end{aligned} \tag{7s}$$

Since the seeding wave progresses through m groups and each group has one-fourth the members of the next lower risk group, $m = \ln \frac{N_0}{2} / \ln 4 \simeq 11$, for a total host size of 4 million, or $\frac{3}{2}N_0$. Thus, the maximum number likely infected at the start of the seeding wave is $m(\frac{2}{F} + 1)^2 \simeq 860$. Of course, the out-of-group mixing fraction F is only poorly estimated, and a factor of two larger or smaller value for F implies a range of 270 to 3000 infected before the start of the saturation wave. Although these estimates cover a wide variation, they are upper bounds on the number infected before the start of the saturation wave. As mentioned above, leapfrogging would circumvent the seeding wave and reduce the number infected prior to the start of the saturation wave. Moreover, these upper bounds are not inconsistent with the prediction in the main text that the size of the infected cohort before the start of the saturation wave, namely I_0 in Eq. 24, is small.

This very simple description of the initial spread of infection opens up a number of questions. (1) What is the likely time when the first individual

was infected and, hence, later became the likely first case of AIDS? (2) Is the predicted risk behavior of the early cases of AIDS, inclusive of the seeding wave, consistent with the high mean risk behavior of the early AIDS cases observed by the Centers for Disease Control (CDC)? (3) What is the probability that the whole process of group-to-group progression is circumvented by one high-risk individual becoming infected early in the seeding process? (4) Is the seeding process consistent with our perception that all major demographic groups participated in a near simultaneous start, that is, synchronization of the saturation wave?

Infection Time. We would like to associate a real time with the time step t_k of Eq. 4s and then take the sum $\sum_1^m t_k$ as the total, or maximum likely, time of the seeding wave. This then becomes the *maximum* time prior to 1979.2 that the first person in the United States was likely to have been infected.

In the seeding-wave process, the growth rate of any given group is $(1 - F)\alpha r_k$, where the factor $(1 - F)$ recognizes that out-of-group mixing is not balanced by equal and opposite in-group mixing. We now use the current growth rate of the group at the front of the saturation wave to calibrate $(1-F)\alpha$. In this way, we derive a very rough estimate for the maximum time of the seeding wave.

Figures 2 and 3 of the main article indicate that, at the time 1988.2, the homosexual fraction was approximately 65 per cent of our estimated one million infected, which is 650,000 infected, or $\frac{1}{6}$th of our estimate of the total number of active homosexual population of 4 million. This estimate places the presently infected population partly in group 1 with all higher groups near saturation. The total population already infected in the higher risk groups is $\frac{4}{3}N_2$, or roughly 330,000 (Eq. 2s). Thus, about the same number must be infected in group 1 so that the total is 650,000.

It has required 9 years for the seeded fraction of 81 individuals in group 1 to grow to 320,000, which gives a growth rate of $(1 - F)\alpha = \frac{1}{9} \ln \frac{320000}{81} = 0.92$ per year or a doubling time of 0.75 years. Thus, the apparent growth rate for the total epidemic, which must be averaged over both group 1 and all higher risk groups—groups that, by now, are almost saturated, gives a doubling time that is roughly twice as large, or 1.5 years. This doubling time is to be compared to the present doubling time for infection predicted by our saturation-wave model, which, at $t + 2 = 9$ years, is $(\frac{1}{I}dI/dt)^{-1} \ln 2 = 0.69t/2 = 3$ years. The three-year doubling time corresponds to a two-year doubling time for AIDS, in agreement with present CDC estimates of 1.75 years. Thus, our saturation-wave model is consistent with the CDC data but inconsistent with the simple seeding-wave growth by a factor of two. One source of discrepancy is our incomplete treatment of the effects of out-of-group mixing. We therefore estimate that the growth rate in group 1 is bounded by a doubling time of 0.75 to 1.5 years.

In Eq. 2s we have neglected group 0 $(0 < r < 1)$ with 3.3 million individuals. The first individual infected is equally likely to be in group 0 or 1 because the average value of $N(r)r$ is approximately the same for both groups. We neglected group 0 to simplify the seeding-wave calculation, but since our estimates for the doubling time are too short, we must now recognize that the initial infected individual most likely had a lower mean risk than group 1 and that the mean growth rate is between the growth rate of two groups. As a rough approximation, let us say that the mean growth rate is lower by a factor of $1/\sqrt{2}$. Then the doubling time of the combined group average will be $0.75\sqrt{2}$ to $1.5\sqrt{2}$ years, or 1.1 to 2.2 years. This then becomes a rough estimate of the doubling time of the seeding wave.

First Infection. The date for the beginning of the saturation wave or power-law (t^2) growth of infection was 1979.2 (Eq. 24). But the seeding-wave model suggests that the first infection in the United States may have occurred $\ln((\frac{2}{F} + 1)^2)/\ln 2 \simeq 6$ doubling times earlier, or 7 to 14 years earlier. The date of the first infection thus may fall somewhere between 1972 to 1965, earlier than has previously been estimated. Thus, the singular case of a teenage boy in St. Louis who has now been identified as having died of AIDS in 1969 is consistent with our seeding-wave picture if he was infected up to five years before developing AIDS. The existence of this case of AIDS in 1969 implies a slow growth of the number infected before the start of the saturation wave.

Mean Risk Behavior. We wish to confirm that our model of the seeding wave, which starts in relatively low-risk groups, is consistent with the CDC observation that most early cases of AIDS were among high-risk individuals. The mean risk behavior of those developing AIDS can be calculated using a convolution integral similar in structure to Eq. 26, but one emphasizing risk rather than time since infection. However, here we are really interested in risk behavior versus time and the absolute number of cases of AIDS, because it was the occurrence in 1981 of roughly 50 AIDS cases in a relatively short period of time (approximately 6 months) that caused the recognition of an epidemic.

We next must define high- and low-risk behavior in terms of our seeding-wave model. The new-partner rate of the homosexual population in London SDT clinics (Fig. 5 in the main text) has a mean of roughly 24 partners per year.

We associate this new-partner rate with group 1 of the seeding-wave model. Group 2 would then have a mean rate of 48 new partners per year—well within the CDC definition of extremely high risk behavior. Thus, moderate or low risk behavior is restricted to groups 1 and 0 with doubling times of 1.1 to 2.2 years and 0.8 to 1.6 years, respectively. By 1979 these two groups would each have infected $(\frac{2}{F}+1)^2 \simeq 81$ individuals. Two years later, in 1981, the combined groups would be producing AIDS cases at a rate of 6 per cent per year, that is, $0.06 \times 2 \times 81$, or 10 cases per year. The total cases for 1981 was several hundred, so these 10 additional moderate-risk cases would, by comparison, be negligible. Thus, we believe that the seeding-wave model is consistent with the CDC observation that high-risk behavior was strongly correlated with the AIDS cases at the start of the epidemic.

Bypassing the Seeding Wave. Of course, this slow growth for 7 to 14 years in group 1 could have been bypassed by one member of any group with $r \geq 2$ becoming infected at the beginning. The probability of this happening per infection in group $k-1$ is, for each group, proportional to $N_k r_k \propto F/(2r_k)$ per group, discounting mixing biases. Therefore, the probability that at least one member of higher risk becomes infected, exclusive of the seeding wave, becomes

$$P = \frac{F}{2}\sum_{2}^{m}\frac{1}{r_k} \cong \frac{F}{2}. \qquad (8s)$$

That is, when one member of group 2 becomes infected, it is equally likely that a member of any higher risk group will become infected, and then the remaining time to the start of the saturation wave becomes negligibly small. This effect would reduce the time for the start of the saturation wave by a factor of 1/2 or less, or just the time to infect one member of group 2, which is within the error of our estimates. On the other hand, if we wish to preserve this factor of 2, we must require that F is a function of r_k/r_{k+n}. A bias function as weak as $F \rightarrow F/\ln(r_{k+n}/r_k)$ guarantees that the roughly 100 out-of-group infections likely to occur during the course of the seeding wave have a small probability of being in the highest risk group ($k+1 \leq m$). Otherwise, infection will leapfrog to reach saturation in one-half the time.

The seeding time would likewise be shorter if the infectious source (in another country, for instance) grew rapidly enough to cause many infections in group 1 and, hence, at least one infection in higher risk groups. There is also the possibility that the infection started and died out several times in groups 0 and 1 before starting the seeding wave. This possibility is equivalent to saying that the net reproductive rate of the disease is very close to unity in these low-risk groups, that is, that a given infected individual infects only one other in the mean time of 10 years to AIDS and death. Because of the arguments in the main text concerning the probability of infection per sexual contact and the equivalence of new-partner rate and contact frequency, we believe the net reproductive rate in the homosexual population was large, and thus the seeding wave started with the first infection.

Synchronization of Risk Populations. We ask if either the slow seeding wave or the singular high-risk initial case of infection makes any difference to the saturation-wave model. For sexual preference and race (Figs. 3 and 4 in the main text) as well as regional and age populations (not shown), the cube root of the cumulative number of cases is nearly linear for $t \geq 1982.5$. These curves extrapolate to zero at approximately the same time with a maximum delay of half a year. This result means that all these subpopulations had to be seeded with at least one high-risk member infected within this time interval. The number infected after half a year (using Eq. 24) becomes roughly 2000 or 3000, all within high-risk categories and with or without the seeding wave of initial infection. We then ask what the probability is that any population selected does not have one member within the 2000 to 3000 initially infected high-risk groups. This depends upon the social isolation, but for a subdivision that creates 10 or more categories, no one population is likely to have less than 100 to 150 members in its high-risk group. Thus, isolation would have had to been very strong, such that none were infected. The observed synchronization of the subpopulations seems reasonable and is independent of whether a seeding wave or single high-risk infection started the epidemic.

In summary, we have described a plausible process by which the initial infection of HIV spread in the various risk populations of the United States. Initially, an average individual was infected from sources unknown, but the infection then grew in a peer group until the number infected and the probability of out-of-group mixing caused the infection to jump to a higher risk group. In this fashion, a seeding wave of infection steadily climbed to the highest-risk individuals. The rapid growth among these highest-risk individuals caused all of them to be rapidly infected, resulting in the start of saturation-wave growth for the whole population. The total number infected in the initial seeding wave is strongly dependent upon the out-of-group mixing fraction, but reasonable estimates indicate that the number infected by the seeding wave would be small enough, less than several thousand, to leave the later saturation-wave growth intact. The earliest known case of AIDS in the U.S. in 1969 is consistent with this picture. ■